Rating Statistical Reasoning

Brian Junker and Beth Whiteman

9/10/2020

Table of Contents

Introduction

- We worked with three graduate assessment fellows in 2019-2020:
  - Refine and try out the rubrics for *Statistical Reasoning*
  - Focus on artifacts from 36-200 “Reasoning With Data”.
- This presentation:
  - May 2019 Initial Learning Goals and April 2020 “Final” Rubrics
  - Quantitative analyses
    - Rater agreement, regression analysis, mean ratings
- Possible extension:
  - Relate our results with text analysis of the prompts and artifacts (“Project 2” papers).

Statistical Reasoning - Learning Goals in Draft Curriculum (May 2019)

- Learn the empirical research process including data collection and experimental design methods
- Develop and use methods for summarizing and evaluating different types of data
- Learn and apply the basic concepts of probability and hypothesis tests
- Develop skills in the application of statistical methods to problems in the humanities, social sciences, as well as physical sciences, including interpretation and communication of results
- Design research questions and correctly utilize appropriate statistical methods to draw conclusions; disseminate the related work in written or graphic form

Statistical Reasoning - Revised Goals/Rubrics for Rating (April 2020)

- Divided into two parts: *Study Design*, and *Inference/Reporting*
- *Part 1: Study Design*: The student…

  – *RsrchQ:* Generates or critiques an empirical research question.

  – Given an empirical research question, designs a study to convincingly answer the question. **Not applicable to 36-200**

  – *CritDes:* Given an empirical research question, critiques to what extent a study convincingly answers the question.

Statistical Reasoning - Revised Goals/Rubrics for Rating (April 2020)

- *Part 2: Inference/Reporting*: The student…
  - *InitEDA:* Appropriately describes data & provides initial EDA.
  - *SelMeth:* Selects appropriate methods to analyze a dataset.
  - Implements the selected analytic method(s).
    Not rated for 36-200: Implemented via ISLE
  - *InterpRes:* Interprets results of the selected method(s).
  - *VisOrg:* Communicates Effectively - Visual Organization, Coherence (e.g., charts, tables)
  - *TxtOrg:* Communicates Effectively - Text Organization, Coherence

Rating Scales for All Rubrics:

- NA: This rubric not applicable (almost never happened in our rating trial)
- 1: Student fails to generate relevant evidence
- 2: Student generates evidence with significant flaws
- 3: Student generates competant evidence: only minor, or no, flaws
- 4: Student generates outstanding evidence: comprehensive and sophisticated

Much scaffolding in the rubrics was needed to attain fair rater agreement among assessment fellows from different disciplines.

Rating Study design

- 91 Artifacts (Project 2 papers) sampled from 364 students in 36-200, Spring 2019 or Fall 2019
  - 13 artifacts seen by all three raters (assessment fellows)
  - 78 seen by a single rater only (~26 artifacts for each rater)
- Breakdown by Semester and Sex:

	% S19	% F19	% Female	% Male
Population (n=364)	32	68	57	43
Sample (n=91)	31	69	55	44

(One student did not report their sex.)

Rater Performance on 13 Artifacts Seen By All Raters – Summary

- **Mean rating ranged from 1.67 to 2.62 (3 = Competent)**
- **Intraclass-correlations (ICCs) & Cohen's Kappas were poor to moderate**
- **Kappa and Percent Exact Agreement measures suggest**
  - Rater 2 may not have been well calibrated with the other two on
    - *Stating a Research Question*
    - *Critiquing Study Design*
    - *Clarity/Organization of Text*
  - Rater 3 may not have been well calibrated with the other two on
    - *Selecting an Analysis Method*
- **Percent Agreement Within +/-1 was generally high, except on *Interpreting Results*, where Rater 2 again differed from the other two raters**

Rater Performance on 13 Artifacts Seen By All Raters – Details

	ICC	k12	k13	k23	a12	a13	a23	b12	b13	b23	m	sd
RsrchQ	0.16	-0.32	0.57	-0.20	15	85	23	92	100	100	2.28	0.56
CritDes	0.55	0.09	0.57	0.12	38	77	38	92	100	100	1.67	0.74
InitEDA	0.47	0.51	0.40	0.84	77	69	92	100	100	100	2.62	0.54
SelMeth	0.49	0.76	0.03	0.12	92	31	38	100	100	100	1.95	0.60
InterpRes	0.20	0.24	0.21	0.18	46	62	38	62	100	62	2.56	0.75
VisOrg	0.57	0.42	0.67	0.71	69	85	85	100	100	100	2.21	0.57
TxtOrg	0.11	0.15	0.45	0.02	46	69	38	92	92	100	2.54	0.64

ICC from random intercept model (0.50–0.75: moderate)

k12, k13, k23 = Pairwise Cohen's Kappa (0.41–0.60: moderate)

a12, a13, a23 = Pairwise Percent Perfect Agreement

b12, b13, b23 = Pairwise Percent Agreement +/-1

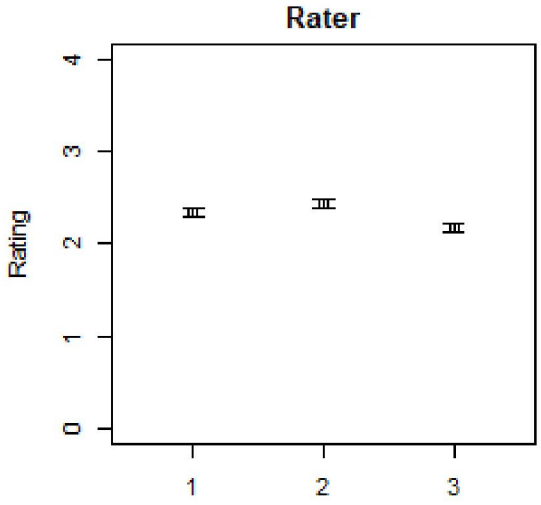
m = sample average

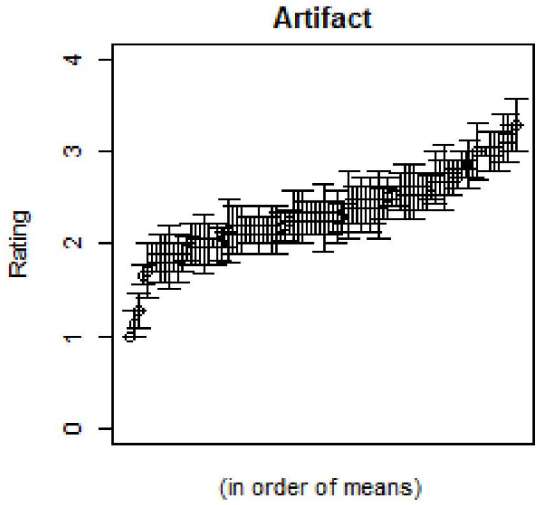
sd = sample standard deviation

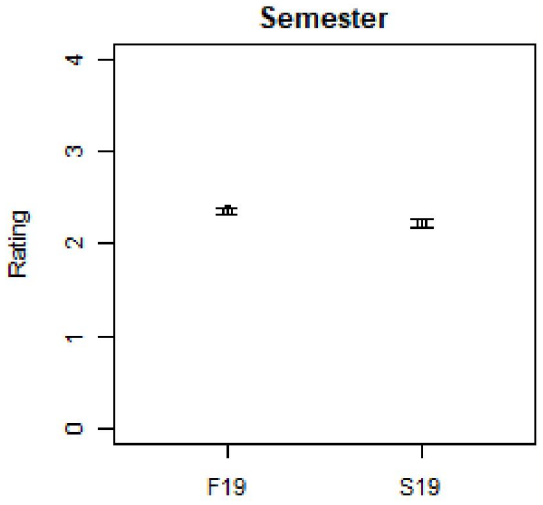
Regression Analyses Using Full Data Set

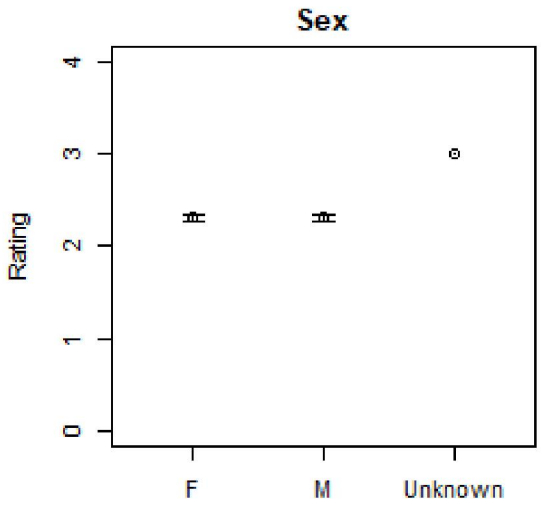
- **We augmented the random intercept model for each rubric with fixed effects for**
  - *Semester*
  - *Sex*
  - whether the artifact was one of the 13 *Repeated* for all raters
  - interaction between *Rater* and *Rubric*
- **Results Using all 91 Artifacts**
  - ICC's were moderately higher in most cases
  - For all but one rubric, very little evidence in favor of keeping fixed effects
  - Ratings for *Selecting Method of Analysis* were approximately 0.32 higher ($p \sim 0.002$) for Fall 2019 artifacts than for Spring 2019
  - Strong evidence ($p \sim 0.000\ldots$) for keeping *Rater* by *Rubric* interaction
- **Regression analyses details omitted from this presentation**

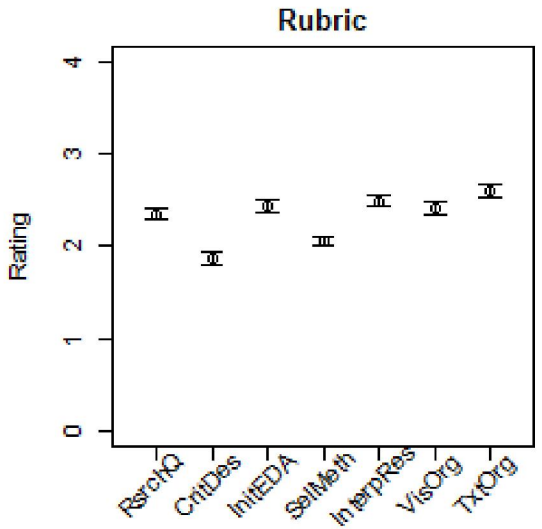
$\overline{X} \pm 2 \cdot SE$ Intervals, Full Data Set

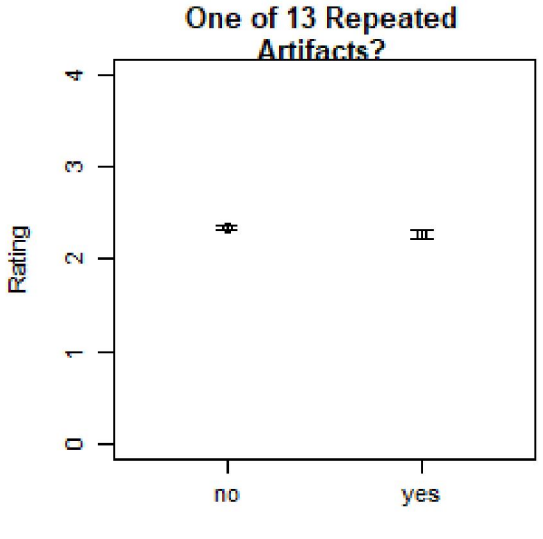




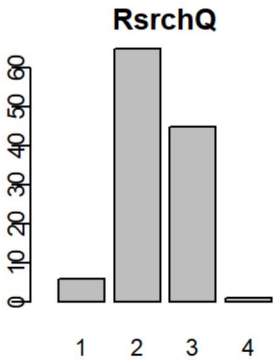


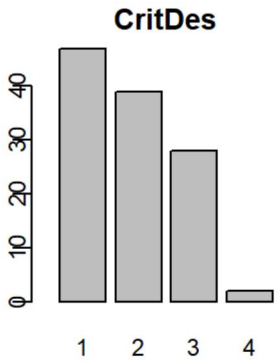


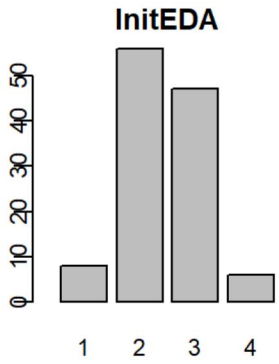


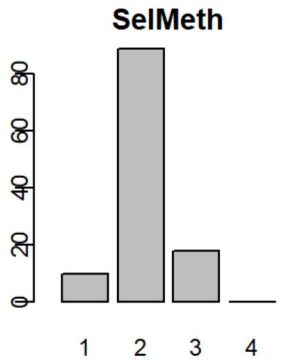


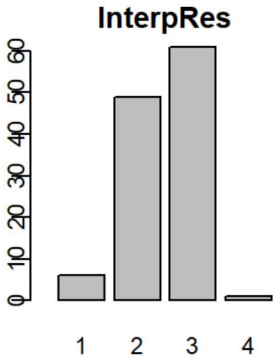
Rating Category Usage, Full Data Set

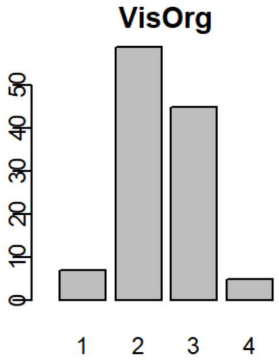


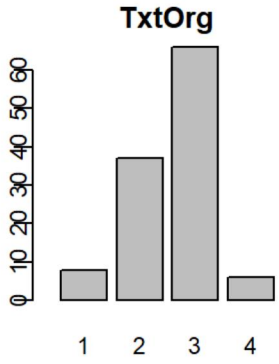












	1	2	3	4
RsrchQ	6	65	45	1
CritDes	47	39	28	2
InitEDA	8	56	47	6
SelMeth	10	89	18	0
InterpRes	6	49	61	1
VisOrg	7	59	45	5
TxtOrg	8	37	66	6

Discussion

- **We developed nine rubrics to measure the five learning goals in the draft curriculum**
  - Two rubrics could not be evaluated using 36-200 data
  - Much scaffolding was required to obtain fair rater agreement
  - *Category 3 (Competent) functions as a virtual ceiling*
- **ICC's and Kappas generally poor to moderate**
  - Some evidence that the raters were not yet well calibrated with one another
  - Percent Agreement Within +/-1 was generally very high, so raters tended to disagree about adjacent rating categories
  - *Probably adequate for averaging ratings across students, not for individual student ratings*
- **Regression analysis** showed a strong Fall/Spring effect for *Selecting Method* and strong evidence for a *Rater* by *Rubric* interaction.
- **Overall mean ratings** confirm the Fall/Spring effect, suggest that Rater 3 may have been harshest, and suggest students do more poorly on *Critiquing Design* and *Selecting Method* than other rubrics.
- **Possible extension:** Relate our results to text analysis of prompts and artifacts.